XAI for
Dummies

User StudiesIII

Ulrike Kuhl

Bielefeld University, Bielefeld, Germany

Mark T. Keane

University College Dublin, Dublin, Ireland

More Complex Designs

- We looked at simple 2-group design and the steps you need to go through
- Then we consider a more complex 3-group design

Control: No-Explanation

Experimental: CF-Explanation

Control:	Experimental#1: XP-method-1	Experimental#2: XP-method-2

- Here we will consider
  - sample size estimation and power
  - aspects of more complex designs (between & within variables)

Six Steps to Heaven…

- STEP 1 Motivation
- STEP 2 Design
- STEP 3 Materials & Procedure
- STEP 4 Piloting
- STEP 5 Data Collection & Analysis
- STEP 6 Results

Single Variable Designs…

(Between- or Within-Participants)

STEP 2

Design

Design-A

Control:
No-Explanation

Experimental:
CF-Explanation

Design-B

Control:	Experimental #1: XP-method-1	Experimental #2: XP-method-2

These are all ***between-participant*** designs, with ***one variable*** (Group), with either 2 or 3 separate groups…

Design-A is a *between-participant design* with one variable, Group, which has 2 levels (Control v Experimental)

Design-B is a *between-participant-design* with one variable, Group, but has 3 levels (Control v Expt#1 v Expt#2)

But, we can also do ***within***-*participant* designs when we give the ***same*** *group* of people ***different*** *treatments* in phases of the experiment (eg before/after tests)

STEP 2

Design

Design-C

Before	After

Design-D

XP1	XP2	XP3

Here are two *within-participant* designs, one variable, Treatment, all with a single group of people

Design-C is a *within-participant design* with one variable, Treatment, which has 2 levels (Before v After)

Design-D is a *within-participant design* with one variable, Treatment, with 3 levels (XP1 v XP2 v XP3)

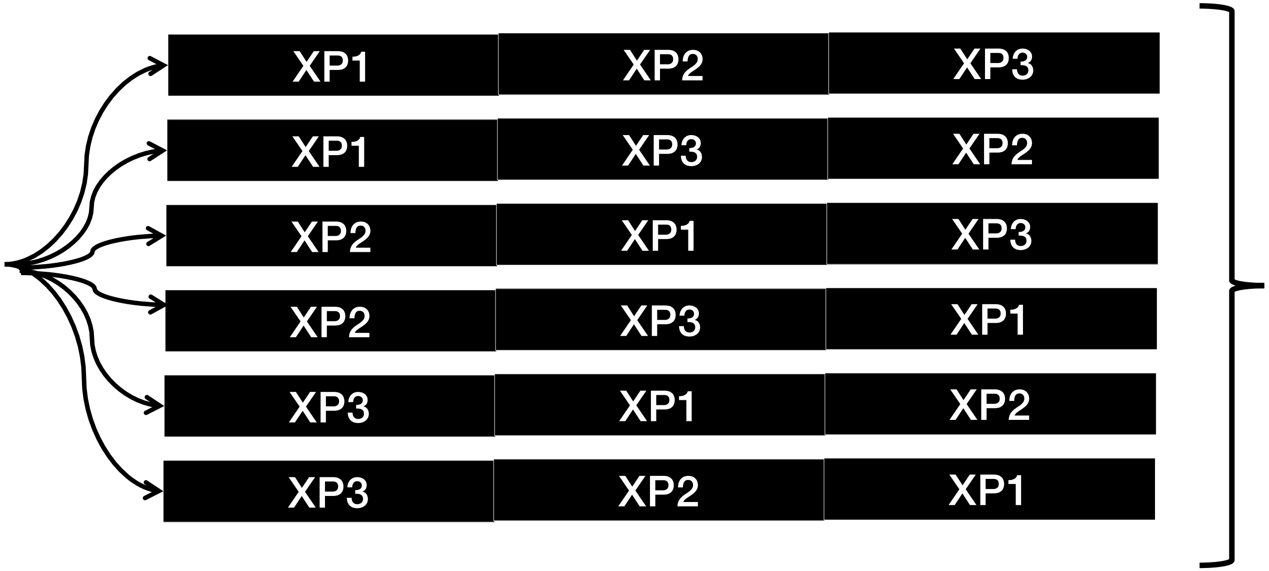
STEP 3

Materials & Procedure

Design D: Presents New Problem?

XP1	XP2	XP3

Order Effects can contaminate this test of the methods; seeing XP1 could affect what people do in XP2 and so on…



XP1	XP2	XP3
XP1	XP3	XP2
XP2	XP1	XP3
XP2	XP3	XP1
XP3	XP1	XP2
XP3	XP2	XP1

XP-Order has 6 levels (o1,o2,o3, o4,o5,o6)

Counterbalance XP-order to check/control XP-Order !

Multi-Variable Designs…

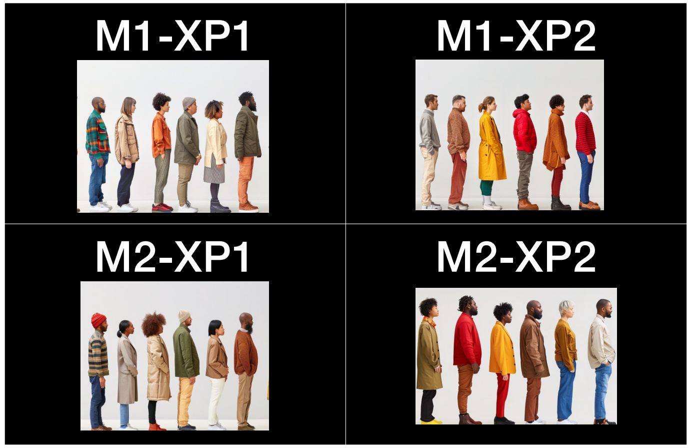
(Between- or Within-Participants)

STEP 2

Design

Design-E

2 x 2



This is a ***between-participant*** design, with ***two variables*** (Material, Method) involving 4 separate groups…

Material has 2 levels (dog-images v cat-images)

Method also has 2 levels (XP1 v XP2)

So we are crossing the variables to test for interactions !

STEP 2

Design

Design-E

2 x 2

NB : we have an order issue here again !!!

This is a ***between-participant*** design, with ***two variables*** (Material, Method) involving 4 separate groups…

Material has 2 levels (dog-images v cat-images)

Method also has 2 levels (XP1 v XP2)

So we are crossing the variables to test for interactions !

Consider Power…

How Many People Do We Test?

Let’s Consider Power

How Many People Do We Test?

- Most common user-study flaw: **too few** participants
- … but you can also test too many !
- ***Type I Error*** = false positive:
  Detect an effect that is not actually there
  --> might happen if *N* is too high!
- ***Type II Error*** **=** false negative:
  Overlook an effect that is there
  --> low *N*

Power Analysis tells you the optimal N for your design based on your estimate of effect-size for a p-level

Let’s Consider Power

How Many People Do We Test?

STEP 2

Design

- Most common user-study flaw: **too few** participants
- … but you can also test too many !
- ***Type I Error*** = false positive:
  Detect an effect that is not actually there
  --> low *N*

Rule of thumb: More Variables = More People!

- ***Type II Error* =** false negative:
  Overlook an effect that is there
  --> might happen if *N* is too high!

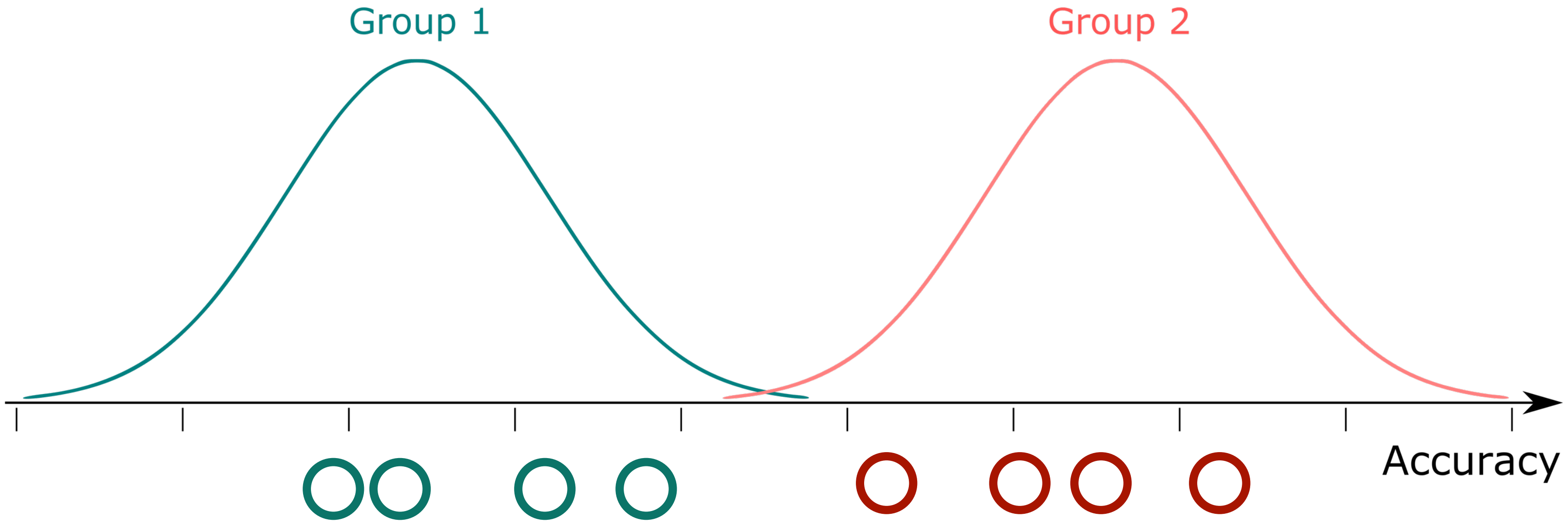
Power Analysis tells you the optimal N for your design based on your estimate of effect-size for a p-level

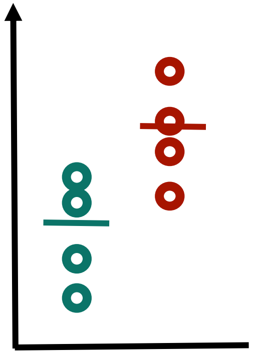
Clear explanations courtesy of Josh Starmer's StatQuest channel: https://www.youtube.com/@statquest

STEP 2

Design

What Is Power?





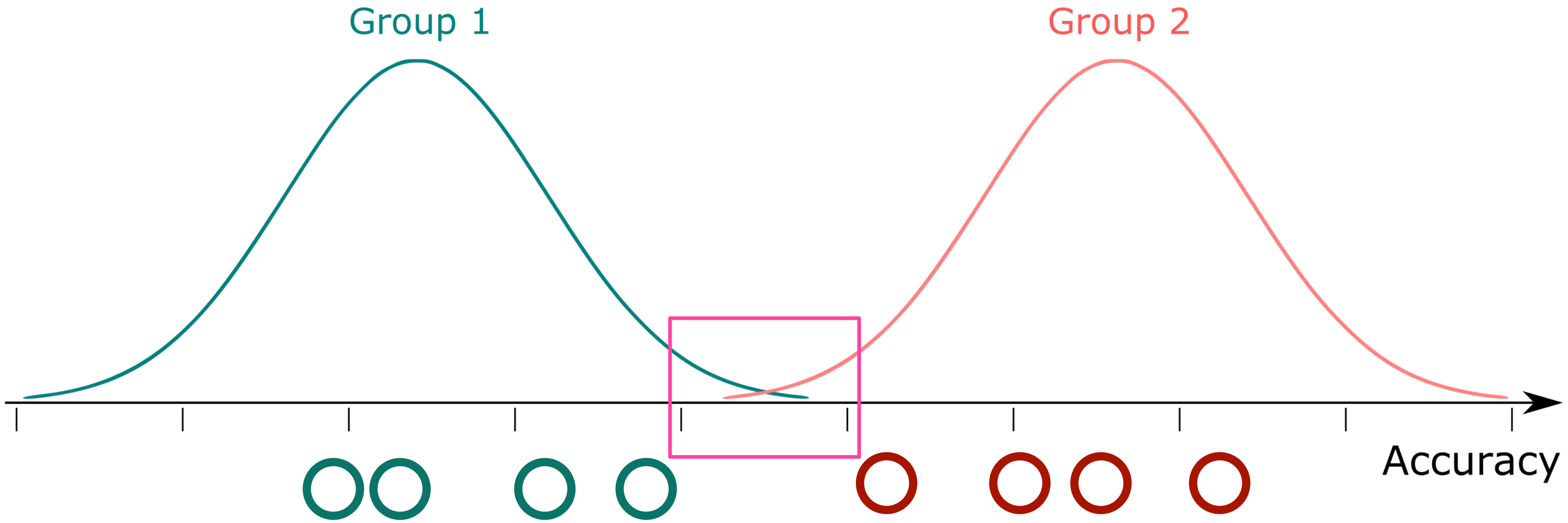
--> test will produce small p-value, correctly indicating the effect!

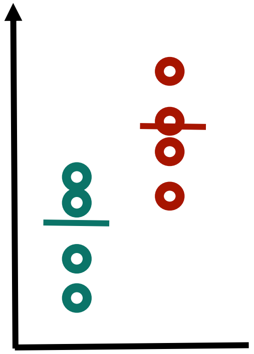
Clear explanations courtesy of Josh Starmer's StatQuest channel: https://www.youtube.com/@statquest

What Is Power?

STEP 2

Design





Small overlap =
high probability to correctly find the effect (even with small samples) = high power

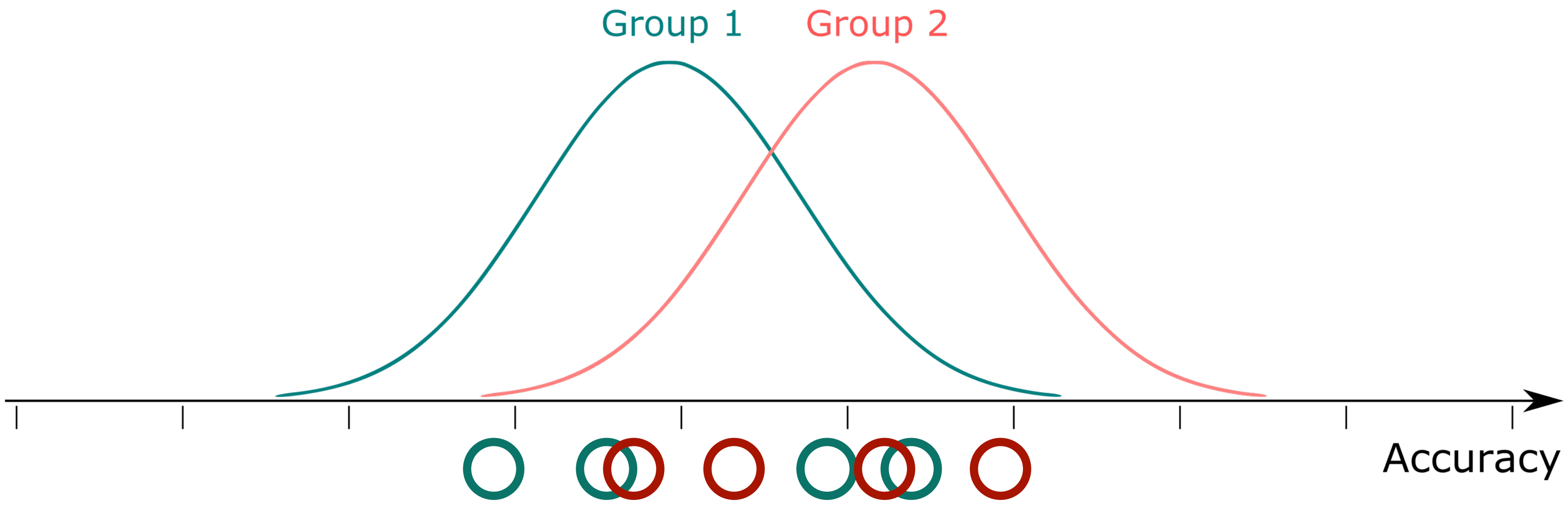
Power is the probability of **correctly finding an existing effect.**

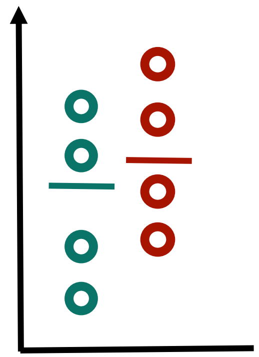
Clear explanations courtesy of Josh Starmer's StatQuest channel: https://www.youtube.com/@statquest

What Is Power?

STEP 2

Design



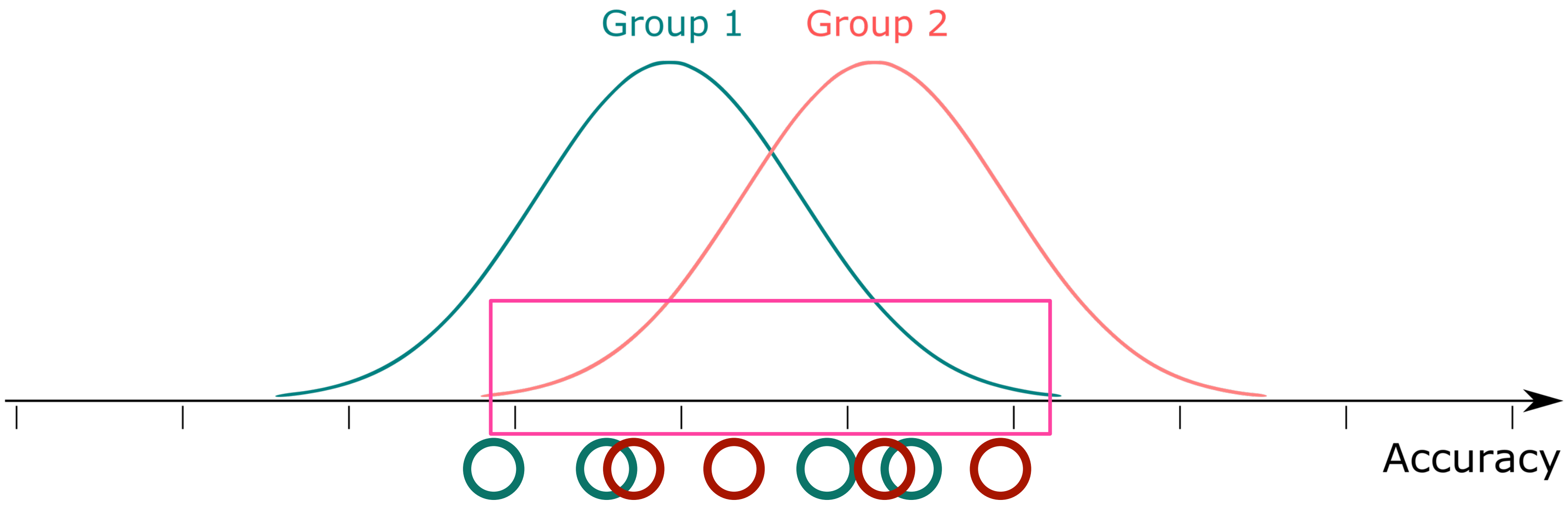


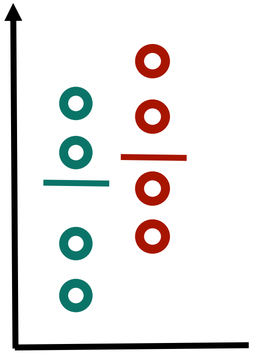
Clear explanations courtesy of Josh Starmer's StatQuest channel: https://www.youtube.com/@statquest

What Is Power?

STEP 2

Design





Big overlap =
low probability to correctly find the effect (esp. with small samples) = low power

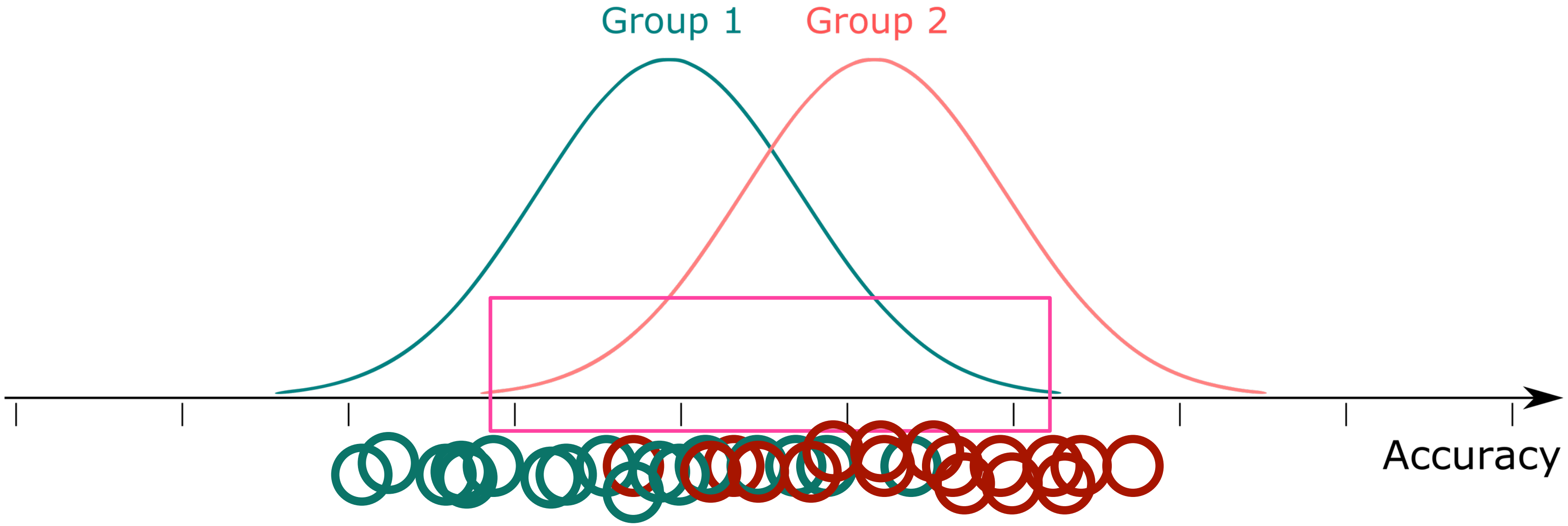
Power is the probability of **correctly finding an existing effect.**

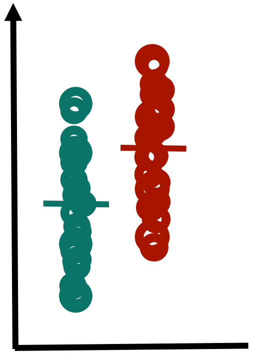
Clear explanations courtesy of Josh Starmer's StatQuest channel: https://www.youtube.com/@statquest

What Is Power?

STEP 2

Design





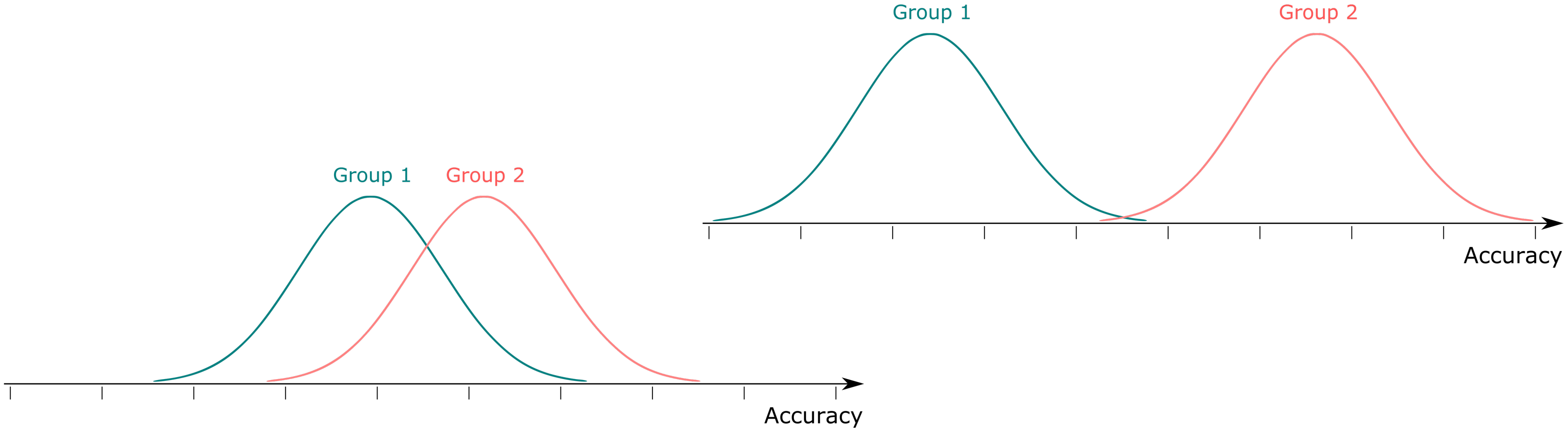
Good news! More samples = more power!

But: Do not just keep samplin guntil the effect is there: <u>p-hacking is evil!</u>

STEP 2

Design

What Is Power?



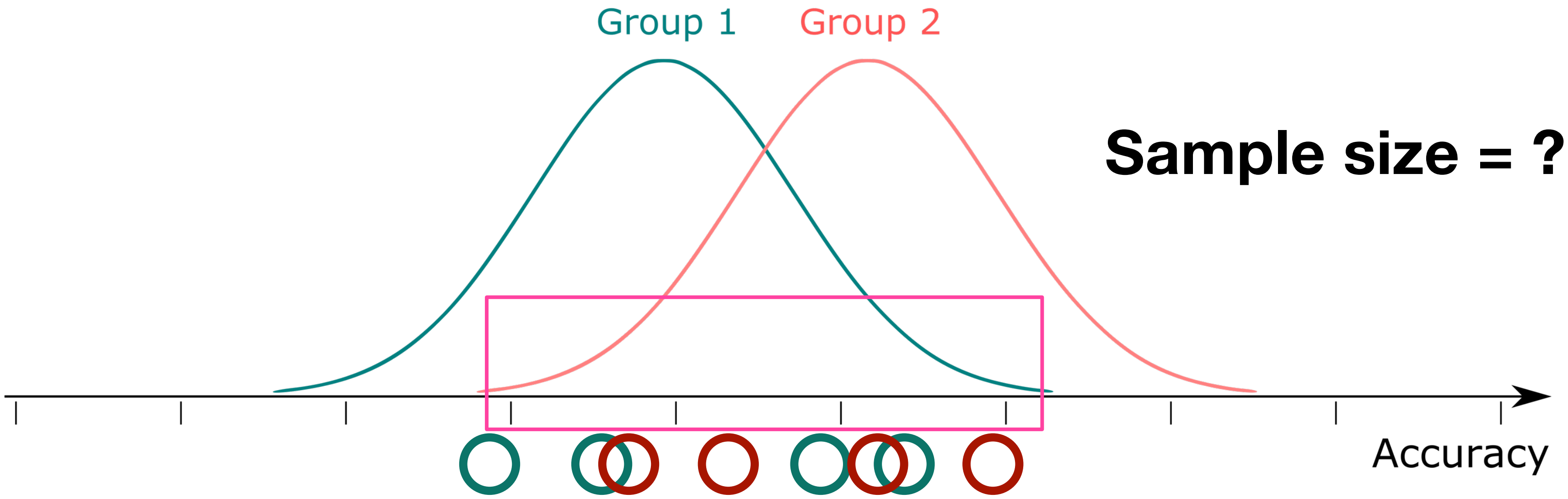
Power Analysis tells us how many measurements we need to collect to have a good amount of power.

If we use the sample size recommended b y the power analysis, we know that we use d enough data to make a decision.

Necessary Decisions

STEP 2

Design



Sample size = ?

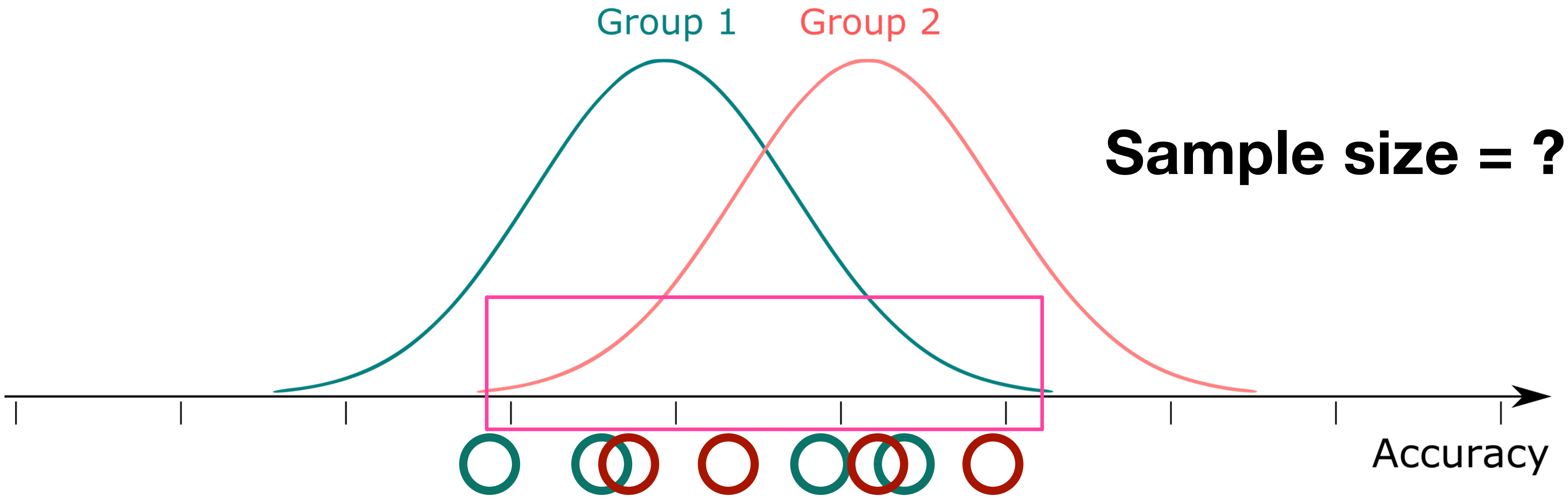
1. How much power do we want? Convention: Power = 0.8
   Meaning: we want an 80% probability of correctly finding an effect.

Clear explanations courtesy of Josh Starmer's StatQuest channel: https://www.youtube.com/@statquest

Necessary Decisions

STEP 2

Design



Sample size = ?

1. How much power do we want? Convention: Power = 0.8 Meaning: we want an 80% probability of correctly finding an effect.

2. What is our significance level α? Convention: *α* = 0.05

3. How much overlap between our distributions do we expect?

Clear explanations courtesy of Josh Starmer's StatQuest channel: https://www.youtube.com/@statquest

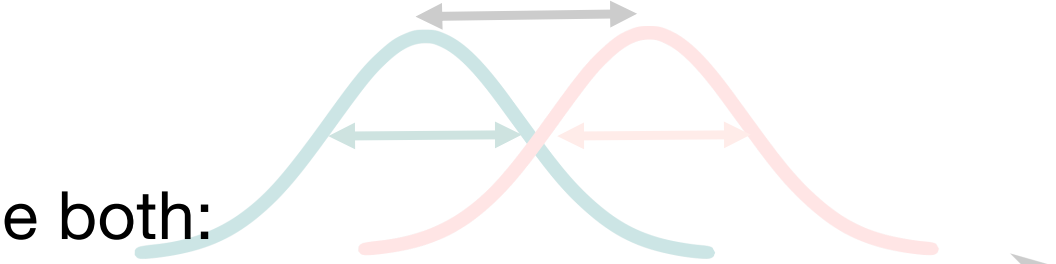
Necessary Decisions

STEP 2

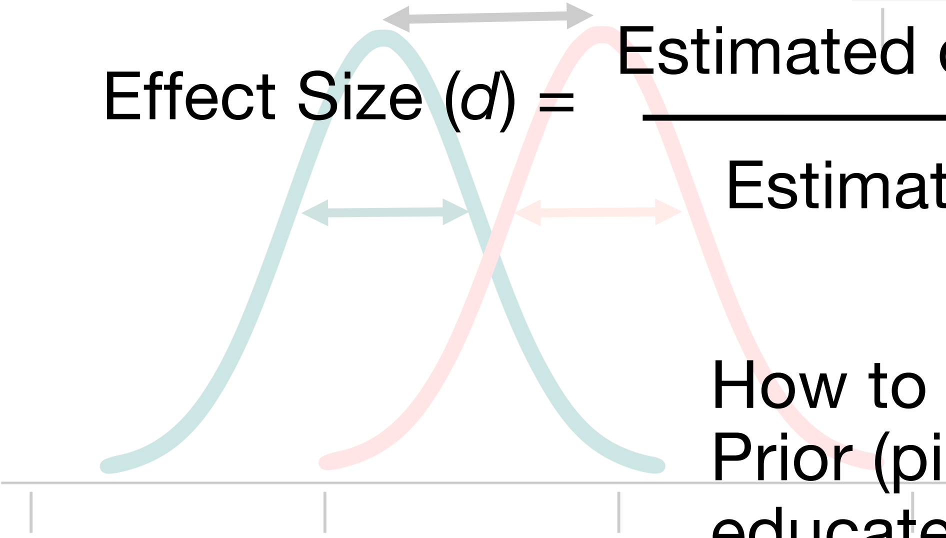
Design

3. How much overlap between our distributions do we expect?

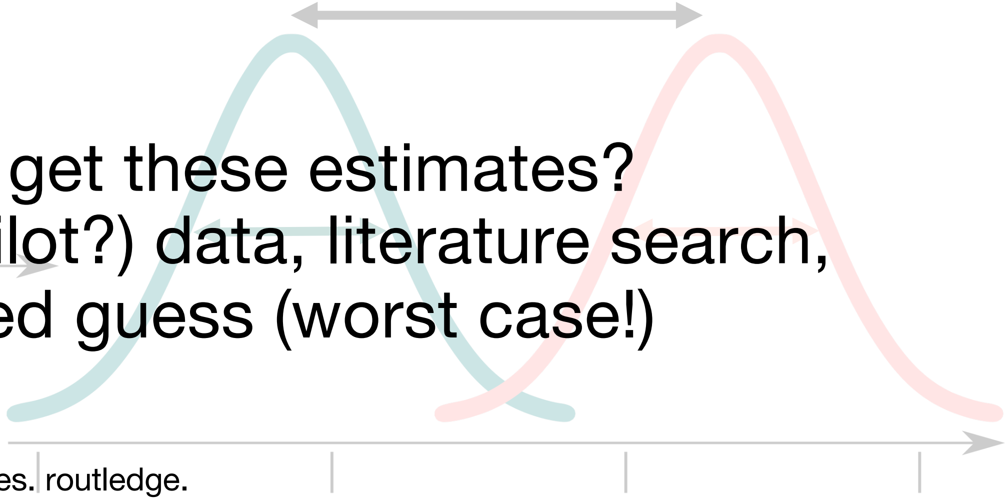
Tricky!



Single metric to combine both:



$$\text{Effect Size } (d) = \frac{\text{Estimated difference between means}}{\text{Estimated standard deviation}}$$



How to get these estimates?
Prior (pilot?) data, literature search,
educated guess (worst case!)

Cohen, J. (2013). Statistical power analysis for the behavioral sciences. routledge.

Sample sizes for different designs: Power = 0.8; $\alpha = 0.05$ and a medium-effect size $d = 0.5$

Design-A

Control:
No-Explanation

Experimental:
CF-Explanation

two conditions using a t-test

N = 128

*64 part. per group

Design-B

Control:	Experimental#1: XP-method-1	Experimental#2: XP-method-2

three conditions using Kruskal-Wallis

N = 72

*24 part. per group

Sample sizes for different designs:
Power = 0.8; $\alpha = 0.05$ and a medium-effect size $d = 0.5$

Design-C

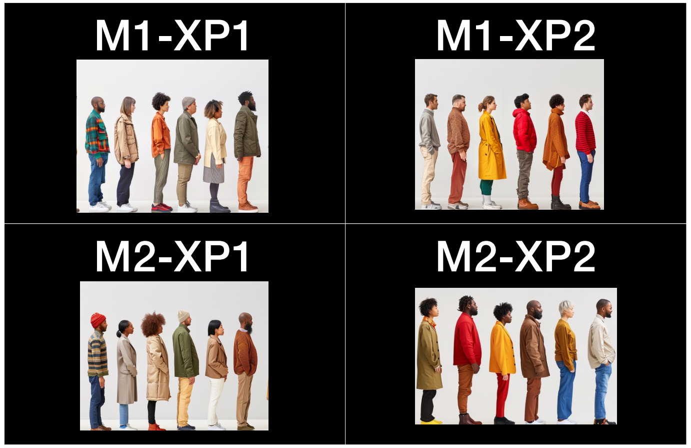
Before | After

two conditions using a paired t-test

N = 34(!)

Design-E

2 x 2



M1-XP1

M1-XP2

M2-XP1

M2-XP2

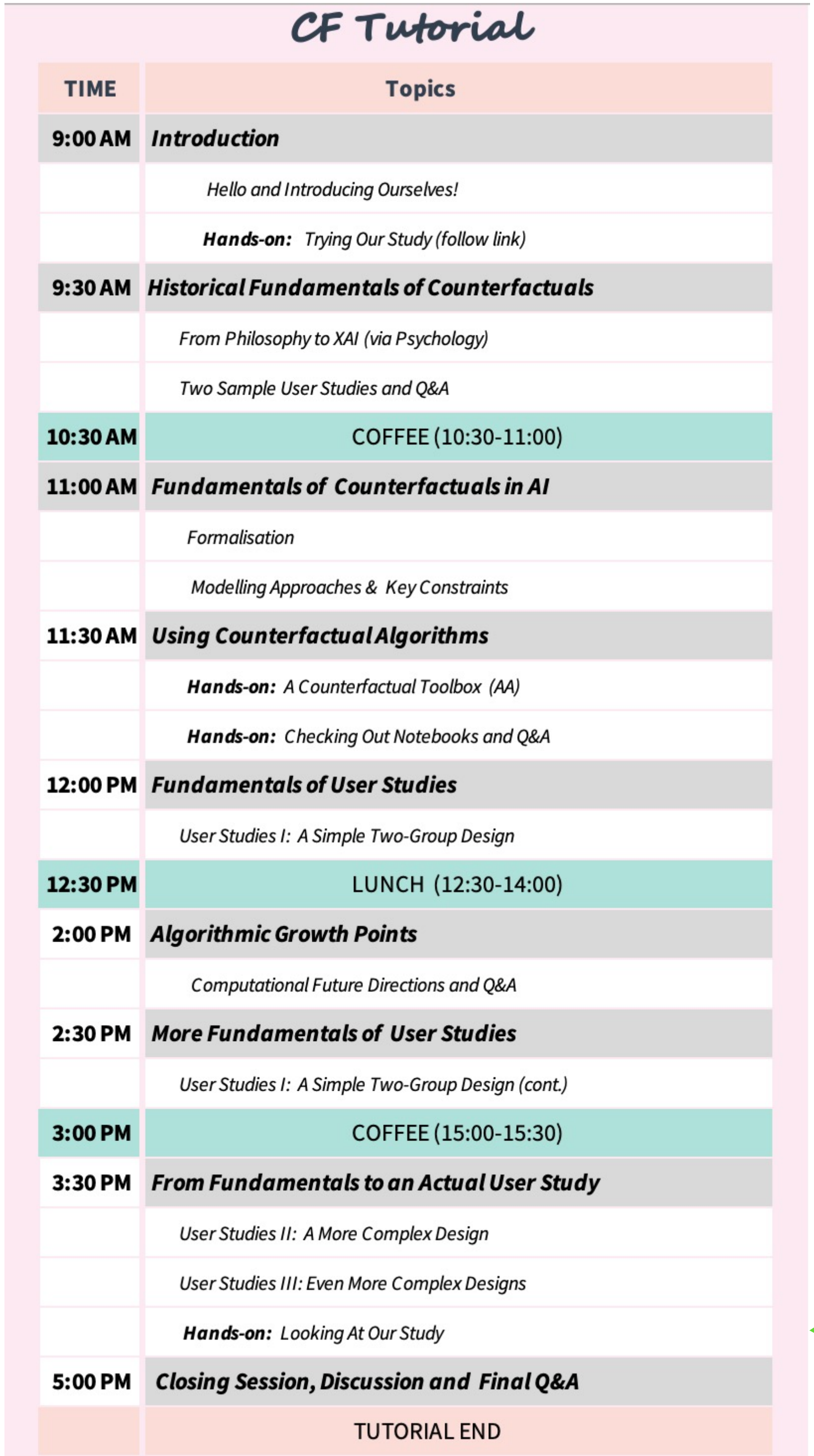
Four independent groups, using a 2x2 ANOVA

N = 180

*45 part. per group

--> Let's come full circle:
What happened in our study

Plan for the Day



CF Tutorial

TIME	Topics
9:00 AM	***Introduction***
	Hello and Introducing Ourselves!
	Hands-on: *Trying Our Study (follow link)*
9:30 AM	***Historical Fundamentals of Counterfactuals***
	From Philosophy to XAI (via Psychology)
	Two Sample User Studies and Q&A
10:30 AM	COFFEE (10:30-11:00)
11:00 AM	***Fundamentals of Counterfactuals in AI***
	Formalisation
	Modelling Approaches & Key Constraints
11:30 AM	***Using Counterfactual Algorithms***
	Hands-on: *A Counterfactual Toolbox (AA)*
	Hands-on: *Checking Out Notebooks and Q&A*
12:00 PM	***Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design
12:30 PM	LUNCH (12:30-14:00)
2:00 PM	***Algorithmic Growth Points***
	Computational Future Directions and Q&A
2:30 PM	***More Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design (cont.)
3:00 PM	COFFEE (15:00-15:30)
3:30 PM	***From Fundamentals to an Actual User Study***
	User Studies II: A More Complex Design
	User Studies III: Even More Complex Designs
	Hands-on: *Looking At Our Study*
5:00 PM	***Closing Session, Discussion and Final Q&A***
	TUTORIAL END

You Are Here!